IBM Data Science Professional Certificate – Applied Capstone Project

Depicting Country-wise population distribution over recommended venues near London

Roudrav Chakraborty

28th July, 2020

1. Introduction

1.1 Background

London, the capital of England and the biggest city in the entire United Kingdom is probably one of the most famous cities in the world. Also known as “The Swinging City” and “Home of The Big Ben”, it surely is a collection of magnificent places. From ‘Buckingham Palace’ and ‘The Big Ben’ to ‘Westminster Abbey’ and ‘The London Eye’, the city is a galore of attraction points.

That being said, there are always recommended spots in a city that attract people of all tastes and get tagged as ‘Must Go-to’ areas. People from all corner of the city tend to check them out and get to explore the various aspects of these venues. The city being host to a diverse collection of people from different nationalities, enables people from all around the globe to experience the areas.

It would be interesting to see how country-wise trend in people going to a particular recommended place has changed over a particular span of time. This data if made available to those places in a visual manner, might allow the respective venues to improve on their qualities by seeing the trend for different countries. This can lead to higher revenues and might allow the businesses to flourish.

1.2 Problem

The count of the people in the city segregated country-wise and coupled with the likes for a recommended place can give a fairly good idea (by assumption of a scale) about the distribution of the population over those venues. Hence this project aims to depict the country-wise distribution of the population over a recommended area and visualize the change (in trend of people from different nationalities visiting) over a period of time, which when used would ensure that opportunities for improvement or new ideas come up for those places.

1.3 Interest

The visualized data, when made public, would be useful to the recommended places as they can analyze the factors behind their success. It would also be helpful for currently under-performing areas to get delve deeper and find out as to what elements impact a nationality’s preference of a place.

2. Data acquisition and cleaning

2.1 Data sources

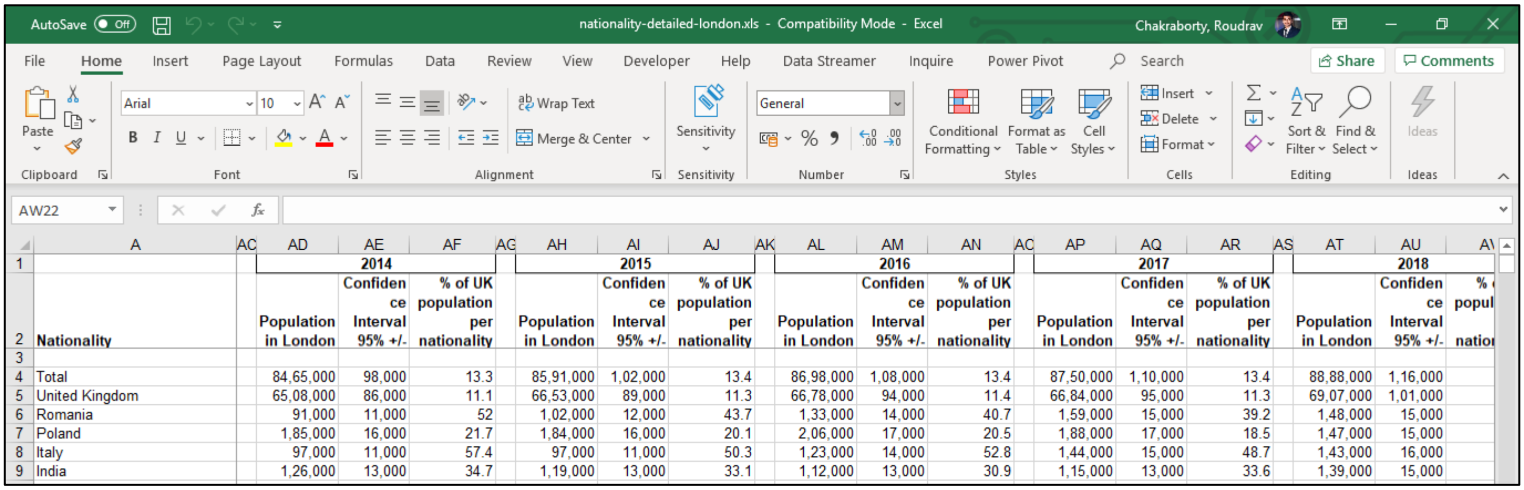
The data was obtained from the London Datastore which contained the **Detailed Nationality** dataset showing the nationality estimates for specific countries outside United Kingdom such as India, United States, Australia, Canada, etc. available for London as a whole. The data set contains the population residing in London per country across 12 years. However, we will be working with the following 5 nationalities for a period of 2014-2018 i.e., 5 years :

- United Kingdom
- India
- United States
- Canada
- Australia

Apart from this, we would be leveraging the Foursquare API endpoints to obtain the **Recommended** places in London at the time of execution. This data would be coupled with the **Likes** per venue to give a brief idea as to how many people would be visiting a recommended place. Upon cross-referencing them with the population dataset, we would depict a pattern of nationalities towards those places.

2.2 Data cleaning

The population data was extracted into pandas dataframe **nat_df**. Firstly, the dataset had multiple columns merged into a single cell, hence they were not labelled properly. The year columns were the only headers that were correctly marked. Below is an example of how the dataset looked like :

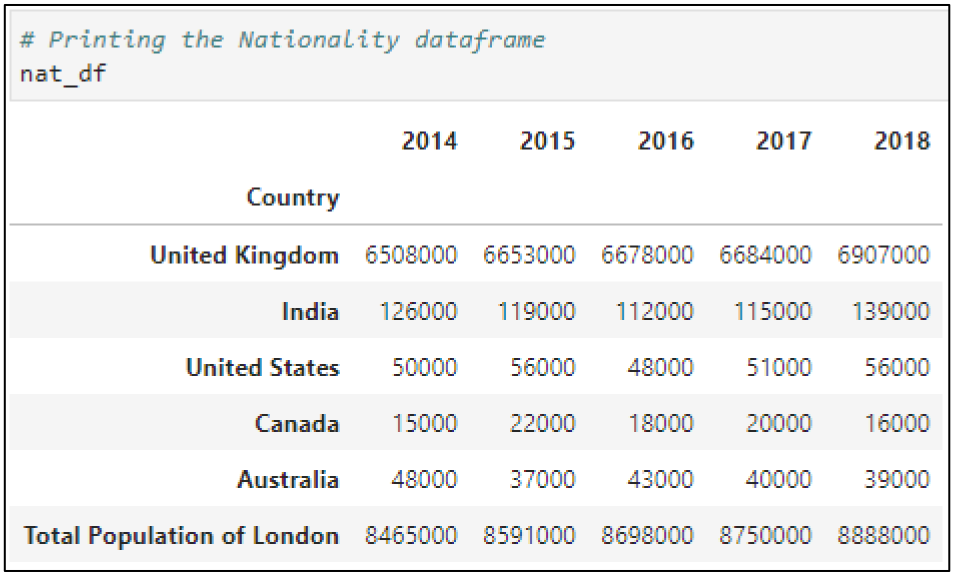


	A	AD	AE	AF	AH	AI	AJ	AL	AM	AN	AP	AQ	AR	AT	AU	AV
1		2014			2015			2016			2017			2018		
2	Nationality	Population in London	Confidence Interval 95% +/-	% of UK population per nationality	Population in London	Confidence Interval 95% +/-	% of UK population per nationality	Population in London	Confidence Interval 95% +/-	% of UK population per nationality	Population in London	Confidence Interval 95% +/-	% of UK population per nationality	Population in London	Confidence Interval 95% +/-	% popul natio
3																
4	Total	84,65,000	98,000	13.3	85,91,000	1,02,000	13.4	86,98,000	1,08,000	13.4	87,50,000	1,10,000	13.4	88,88,000	1,16,000	
5	United Kingdom	65,08,000	86,000	11.1	66,53,000	89,000	11.3	66,78,000	94,000	11.4	66,84,000	95,000	11.3	69,07,000	1,01,000	
6	Romania	91,000	11,000	52	1,02,000	12,000	43.7	1,33,000	14,000	40.7	1,59,000	15,000	39.2	1,48,000	15,000	
7	Poland	1,85,000	16,000	21.7	1,84,000	16,000	20.1	2,06,000	17,000	20.5	1,88,000	17,000	18.5	1,47,000	15,000	
8	Italy	97,000	11,000	57.4	97,000	11,000	50.3	1,23,000	14,000	52.8	1,44,000	15,000	48.7	1,43,000	16,000	
9	India	1,26,000	13,000	34.7	1,19,000	13,000	33.1	1,12,000	13,000	30.9	1,15,000	13,000	33.6	1,39,000	15,000	

Hence, I first generated a list containing the first column (which was the country name) and the 5 consecutive years vis-à-vis 2014 to 2018. From the entire dataset, I extracted the data for only these 6 columns.

Next, due to blank cells at the top of the sheet and the bottom of the sheet, a lot of NaN values were present. So the first 2 records and the last 3 records were dropped from the dataframe to ensure that only the respective county names are present along with the years. Then, I set the 'Country' column as the index of the dataframe so that it becomes easier to slice and data retrieval is simplified.

Here is how the cleaned dataframe looked like :



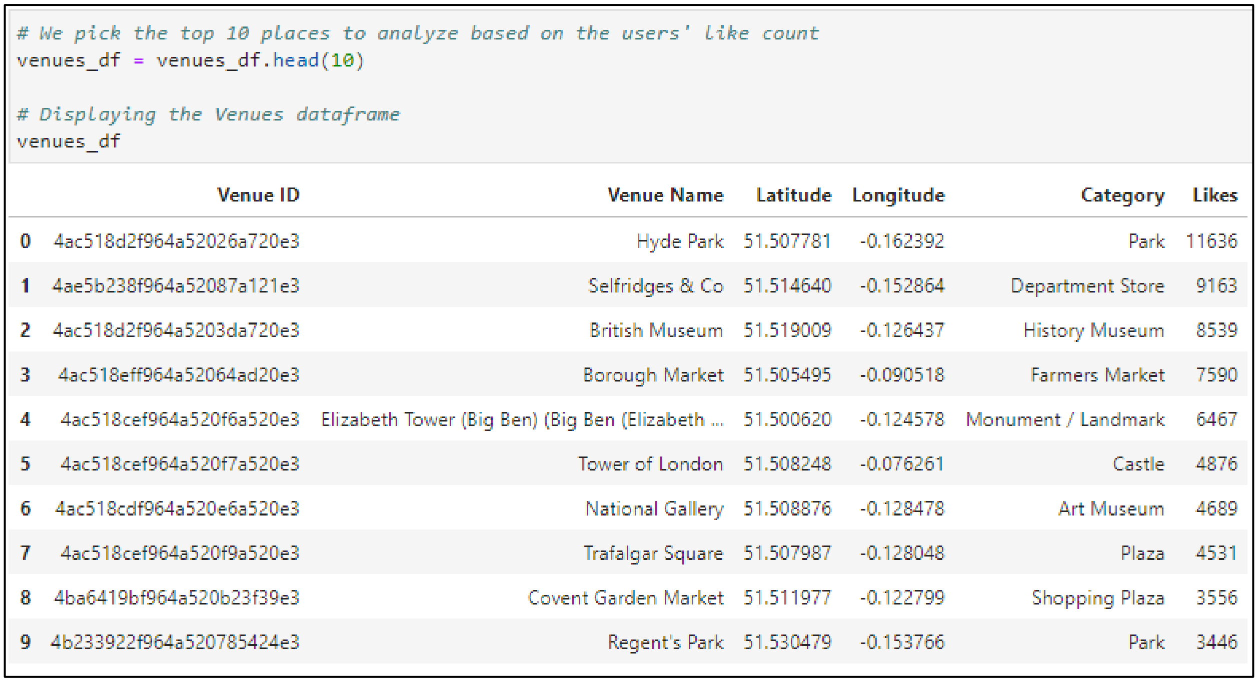
```
# Printing the Nationality dataframe
nat_df
```

	2014	2015	2016	2017	2018
Country					
United Kingdom	6508000	6653000	6678000	6684000	6907000
India	126000	119000	112000	115000	139000
United States	50000	56000	48000	51000	56000
Canada	15000	22000	18000	20000	16000
Australia	48000	37000	43000	40000	39000
Total Population of London	8465000	8591000	8698000	8750000	8888000

Next, I needed the recommended venues within a specific range of the city of London. I defined a custom radius of 50000 meters (50 Kilometers) for extracting the data and fetched up to a limit of 100 venues. The response of the **Foursquare Explore** endpoint was JSON collection which required to be wrangled in order to extract meaningful information. It was the same case for **Foursquare Likes** endpoint, which returned the data in a JSON format.

In both cases, the lists and dictionaries in the JSONs were indexed and the required information was extracted. The dataframe **venues_df** for Venues was created first and then a list containing the Like counts was appended to the dataframe as a column. The dataframe was then sorted in descending order on the basis of the Like count and the top 10 rows (corresponding to the top 10 recommended places) were stored.

The final cleaned dataframe looks as follows :



```
# We pick the top 10 places to analyze based on the users' like count
venues_df = venues_df.head(10)

# Displaying the Venues dataframe
venues_df
```

	Venue ID	Venue Name	Latitude	Longitude	Category	Likes
0	4ac518d2f964a52026a720e3	Hyde Park	51.507781	-0.162392	Park	11636
1	4ae5b238f964a52087a121e3	Selfridges & Co	51.514640	-0.152864	Department Store	9163
2	4ac518d2f964a5203da720e3	British Museum	51.519009	-0.126437	History Museum	8539
3	4ac518eff964a52064ad20e3	Borough Market	51.505495	-0.090518	Farmers Market	7590
4	4ac518cef964a520f6a520e3	Elizabeth Tower (Big Ben) (Big Ben (Elizabeth ...	51.500620	-0.124578	Monument / Landmark	6467
5	4ac518cef964a520f7a520e3	Tower of London	51.508248	-0.076261	Castle	4876
6	4ac518cdf964a520e6a520e3	National Gallery	51.508876	-0.128478	Art Museum	4689
7	4ac518cef964a520f9a520e3	Trafalgar Square	51.507987	-0.128048	Plaza	4531
8	4ba6419bf964a520b23f39e3	Covent Garden Market	51.511977	-0.122799	Shopping Plaza	3556
9	4b233922f964a520785424e3	Regent's Park	51.530479	-0.153766	Park	3446

2.3 Feature selection

For the Nationality dataframe, the data was selected for the following countries :

- United Kingdom
- India
- United States
- Canada
- Australia

The year range was selected as 2014 – 2018 to allow proper non-cluttered visualization. For the Venues dataframe, unnecessary columns such as address, referrals etc. were excluded and only the relevant fields such as Venue ID, Venue Name, Coordinates (Latitude and Longitude) and Category were kept. A list of Likes per venue was added as a column in the Venue dataframe.